Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Clasificación de perfiles académicos mediante técnicas de Machine Learning

Tesis presentada para obtener el grado de:

Licenciatura en Ingeniería en Ciencias de la Computación

Presenta: Jimena Paola Cilia Romero

Director de Tesis: Dr. María Teresa Torrijos Muñoz

Asesor de Tesis: Mtro. Carlos Armando Ríos Acevedo

Puebla, Pue., 22 de noviembre de 2024

Resumen

La educación superior es la principal encargada de formar profesionales competentes en donde desarrollaran un pensamiento crítico, reflexivo y creativo, teniendo un fuerte impacto económico, laboral y social que incrementa el crecimiento profesional, teniendo mejores condiciones de vida. Asimismo, dota a los estudiantes de las competencias necesarias para que respondan a la evolución constante del mercado laboral. Para ello, es importante tener una planta académica sólida y de calidad que ofrezcan una enseñanza efectiva y significativa en los estudiantes, por eso es necesario evaluar y distribuir eficazmente a las personas en áreas específicas según sus cualidades y dominios, con un impacto significativo en la calidad de enseñanza en el nivel superior.

Se aplicaron técnicas y herramientas computacionales para identificar el mejor algoritmo de clasificación que categorizara eficientemente un objeto en una clase definida de acuerdo con las áreas de conocimiento.

En la presente tesis, se explica el proceso de implementación y análisis comparativo de los algoritmos para clasificación codificados en Python, usando técnicas de tratamiento de los datos y Machine Learning, con el objetivo de alcanzar una predicción arriba del 70% con base en los datos disponibles.

Índice general

Capítulo 1

Alcance del proyecto

1.1. Antecedentes del Proyecto

En el estudio llevado a cabo por Karina B. y Roberto S. (2015) se examinaron los factores que influyen sobre la deserción de los estudiantes en la carrera de Ingeniería en Informática de la Universidad Gastón Dachary en Argentina, usando técnicas de clasificación en minería de datos. Su investigación demostró la eficacia de los algoritmos de clasificación para identificar las características que provocan el abandono estudiantil.

Por otro lado, el estudio de Román G. (2011) se centró en evaluar el desempeño de los docentes de la Facultad de Ingeniería de la Universidad Nacional Autónoma de México al final del semestre de acuerdo con una serie de preguntas usando la técnica de árboles de decisión, con el fin de brindar herramientas para la toma de decisiones y tener una mejor calidad en la enseñanza. En su estudio relevó el gran aporte que hace la minería de datos para que se refleje de mejor forma la opinión del alumno sobre los profesores.

1.2. Objetivo general y específicos del proyecto

Objetivo General

Identificar un modelo de clasificación que reconozca eficientemente la pertenencia de los objetos a una clase; esto, mediante técnicas de Machine

Learning que permiten imitar la forma en que los seres humanos aprenden, mejorando gradualmente su precisión.

Los objetos de estudio serán personas que tienen un perfil académico de Ciencias Computacionales y las clases corresponderán a un conjunto de Áreas de Conocimiento definidas por el usuario final.

La identificación del modelo considera la implementación de éste.

La implementación se soporta en el uso de una metodología formal de desarrollo de software con enfoque ágil en el marco de Scrum. Para cubrir el ciclo de vida de desarrollo del proyecto se usarán las herramientas de apoyo: Bonita Soft para el modelado de procesos, G Suite que integra herramientas colaborativas para levantamiento de requerimientos, análisis y diseño, SqlDBM para el diseño de la base de datos, Python para la construcción del modelo, SQLServer como manejador de base de datos y PyTest para la realización de pruebas.

Objetivos específicos

Aplicar técnicas de preprocesamiento de datos como la exploración y conversión de datos para prepararlos para su análisis.

Comparar los resultados de los algoritmos de clasificación implementados para seleccionar la técnica más adecuada que haga asignaciones de la mejor manera.

1.3. Metodología.

Tomando como base a García R. (2015) y Molina R., Honores T., Pedreira S. y Pardo L. (2021), se realizó un análisis comparativo de las características propias

del proyecto, así como de las metodologías para el desarrollo de software y se encontró que la metodología ágil Scrum era la idónea para el desarrollo del proyecto. Las metodologías ágiles para el desarrollo de software buscan proporcionar pequeñas piezas funcionales del software. Esta se caracteriza por realizar entregas en periodos de tiempo cortos y continuas de un software funcional, donde los resultados entregados sean de calidad con respecto al anterior y aporten valor a la organización. Además de tener flexibilidad para adecuarse a las necesidades y cambios que se sugieran en las diferentes etapas del ciclo de vida. Con ello, se obtiene mayor productividad donde se podrán predecir los resultados y minimizar los riesgos asegurando trabajos de calidad.

Existe una gran variedad de marcos ágiles con los que se puede trabajar, se decidió usar Scrum que se basa en el empirismo y el pensamiento Lean. En el cual los conocimientos provienen de la experiencia y la toma de decisión se basa en la información existente, reduciendo el desperdicio, es decir actividades que consumen recursos sin aportar valor al proceso, y enfocándose solo en lo esencial. Asimismo, Scrum combina cuatro eventos formales dentro de un evento contenedor, el Sprint. Funcionan gracias a la implementación de los pilares empíricos de transparencia, inspección y adaptación.

Scrum cuenta con tres artefactos fundamentales y esenciales que refuerzan el empirismo, así como maximizar la transparencia de la información clave logrando que todas las personas que los consulten tengan la misma base de adaptación. Los artefactos son el Product Backlog contiene el objetivo del producto, en este se enlistan de forma ordenada de acuerdo con su prioridad todas las tareas que se van

a realizar durante el desarrollo, con una descripción breve sobre lo que se desea para el producto.

El Sprint Backlog es el objetivo del Sprint, un conjunto de elementos del Product Backlog y un plan de acción para entregar el Incremente, estos responden a las preguntas por qué, qué y cómo. Es un recurso visual en tiempo real del trabajo que se realiza en el Sprint, el cual se actualizará de acuerdo con los nuevos conocimientos se vayan adquiriendo en el equipo de Developers.

Finalmente, el Increment tiene el compromiso de la definición de Terminado, la cual es una descripción formal del estado del Increment, que son elementos del Producto Backlog que cumplen con las medidas de calidad requeridas para el producto. Están se van sumando a un listado de todos los ítems completados durante un Sprint y el valor de todos los incrementos de Sprints pasados.

1.4. Infraestructura.

Para la implementación de código se utiliza Python 3 con ayuda del gestor de base de datos SQL Server para almacenamiento y manipulación del conjunto de datos. Asimismo, se utilizan las herramientas que Excel pone a disposición para el manejo de datos.

CAPÍTULO 2

Algoritmos de clasificación

En este capítulo se presenta una descripción conceptual de la clasificación y posteriormente se aborda en el sentido estricto de las Ciencias Computacionales. Se describen los algoritmos identificados para resolver la problemática planteada y finalmente se menciona la importancia de la toma de decisiones informadas en el marco de las Ciencias Computacionales.

2.1. Clasificación

"*Aquel que trata de comprender el mundo no hace más que clasificarlo*" (Caro-Castro, Carmen, 2010, pp.17).

Bajo esta premisa, la ciencia no es estática, desde Aristóteles y hasta nuestros días, los expertos de todas las áreas de conocimiento han clasificado sus objetos de estudio en función de las características que comparten, buscando que la clasificación considere a todos los objetos conocidos en ese momento y que ninguno de éstos se pueda ubicar en dos categorías diferentes (UNAM, 2013).

La historia de las clasificaciones muestra que éstas sirven para ordenar, establecer diferencias, jerarquías y géneros (González Casanova,1996) y, de acuerdo con Fernández Medina (2012), clasificar es una forma de organizar la información que se puede definir como una actividad en la que, en función de algún criterio, se les asigna una categoría a diferentes objetos, conceptos o seres. En este sentido, "*clasificar se refiere a los mecanismos culturales y cognitivos a través de*

los cuales se obtiene; y las clasificaciones resultantes son las representaciones lingüísticas, mentales y culturales que de ella resulta".

Por otro lado, de acuerdo con los principios básicos de las clasificaciones estadísticas en el ámbito sociodemográfico en México, la clasificación constituye un campo de la lógica, consiste en agrupar objetos con base en sus semejanzas y separarlos por sus diferencias y, una clasificación también puede ser vista como una serie ordenada de divisiones y subdivisiones de objetos en grupos distintos pero relacionados entre sí (INEGI, 2005).

En la ciencia contemporánea se ha extendido el uso de los sistemas de clasificación a las disciplinas particulares y se ha generalizado como una práctica de todas las actividades humanas, considerado la importancia de ordenar y establecer límites, así como de entender, identificar y simplificar el manejo de las múltiples clases de objetos.

De esta forma, la clasificación es trivial cuando las categorías son exactas y bien definidas, lo que implica que las características con las que se cuenta son suficientes para determinar, con certeza, la pertenencia de un objeto a la clase.

En otros casos, cuando la definición de la clase es inexacta y ambigua, provoca que la clasificación ubique a los objetos en la intersección de clases siendo necesario reformular los esquemas de clasificación e incluir nuevas categorías y subcategorías que permitan manejar la exactitud y ambigüedad (González Casanova,1996).

El estudio y la investigación en torno a éstos últimos casos, motivan un sin número de investigaciones que requieren de la clasificación confiable para resolver problemas importantes en distintos contextos como el que se presenta en este trabajo de investigación donde el objetivo es clasificar objetos maximizando la exactitud y minimizando la incertidumbre desde una perspectiva objetiva, crítica y sin sesgo alguno.

2.2. Clasificación como quehacer en las Ciencias Computacionales

El aprendizaje es un procedimiento del cual se toma el conocimiento y se genera como resultado un nuevo conocimiento. En la inteligencia artificial existe un subconjunto que permite a un sistema aprender y mejorar de forma autónoma en función de los datos que consumen, a este proceso se le conoce como aprendizaje automático.

Existen dos tipos principales de modelos de aprendizaje automático. Los modelos de aprendizaje supervisado que se caracterizan por utilizar conjuntos de datos etiquetados para entrenar algoritmos que clasifican datos. En este enfoque, se establece un conjunto de datos de entrenamiento para enseñar a los modelos a producir la salida deseada.

Por otro lado, el aprendizaje no supervisado se produce cuando se proporcionan datos de entrada al algoritmo sin ningún dato de salida etiquetado. Posteriormente, este algoritmo buscará los patrones y relaciones en los datos y entre ellos (Amazon, 2024).

Debemos recodar que la clasificación se puede entender como una tarea de organizar datos o información que se puede definir como una actividad en la que,

en función de algún criterio, se les asigna una categoría a diferentes objetos, conceptos o seres (Fernández Medina, 2012), por ello, los problemas de clasificación se encuentran enmarcados dentro del aprendizaje supervisado del machine learning.

La clasificación, en contexto matemático, es la habilidad para adquirir una función que mapee un elemento de dato de una entre varias clases predefinidas. Las características o variables elegidas dependen del problema de clasificación (Haro Silvia, 2018). Existen varios métodos de clasificación entre ellos encontramos:

K-vecinos más cercanos (KNN)

El método de K-vecinos más cercanos se utiliza para estimar el valor de la función de densidad de probabilidad de la pertenencia de un elemento a una clase basándose en la información proporcionada por un conjunto de prototipos o ejemplos. Es relevante destacar que este método forma parte de una categoría de modelos de aprendizaje llamados “perezosos”, esto implica que guarda únicamente un conjunto de datos de entrenamiento y no requiere una fase de entrenamiento específica, todos los cálculos se realizan cuando se realiza una clasificación o predicción.

En la Figura 1, se muestra un ejemplo del funcionamiento del algoritmo, se presenta un diagrama de dos dimensiones (características x1 y x2) los puntos azules representan los objetos de entrenamiento y el punto amarillo que se muestra con un signo de interrogación, representa el objeto a clasificar.

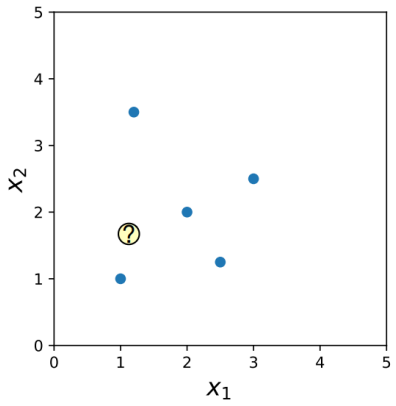


Figura 1. Representación gráfica del conjunto de objetos y su clasificación

En la Figura 2, se muestran las etiquetas de clase y la línea discontinua indicando el vecino más próximo del punto de consulta. La etiqueta de clase predicha es la etiqueta de clase del punto de datos más cercano en el conjunto de entrenamiento, en este caso clase 0.

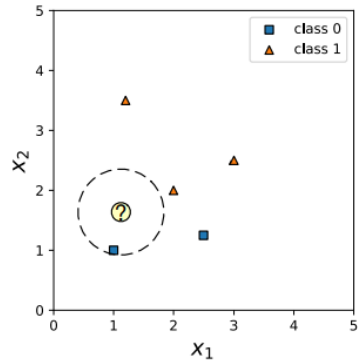


Figura 2. Representación gráfica del vecino más próximo

Arboles de clasificación

De acuerdo con IBM (2023), un árbol de clasificación es un tipo de árbol de decisiones. Calcula la categoría objetivo a pronosticar para cada nodo en el árbol, y, al existir una función que castiga más incluso la distribución de valores objetivos basado en estadísticas de frecuencia de destino y en el número de filas de datos

que corresponden al nodo, cada nodo se divide en dos o más nodos hijos para disminuir el valor de impureza Gini (Figura 3).

Se debe establecer un límite tolerable de impureza Gini logrando que los nodos hijo correspondientes a las categorías de predictores dados se fusionen. Para la división de cada nodo, se selecciona el predictor que mayor disminución aporta al valor de impurezas de Gini.

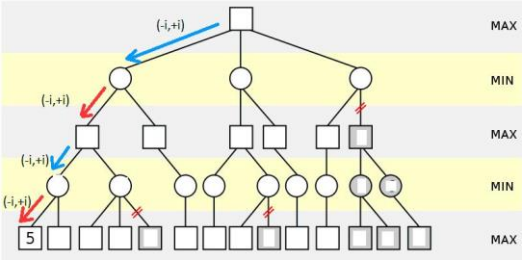


Figura 3. Representación gráfica de un árbol de clasificación

Máquina de Vectores de Soporte

La Máquina de Vectores de Soporte (SVM) es una técnica versátil que se utiliza tanto para la clasificación como para la regresión en el análisis de datos con un gran número de variables predictivas, su principal ventaja es su capacidad de no sobre ajustar los datos de entrenamiento al realizar predicciones de un modelo.

SVM tiene una amplia gama de aplicaciones en diferentes disciplinas, incluyendo la gestión de relaciones con los clientes (CRM por sus siglas en inglés, “Customer Relationship Management”), el reconocimiento de imágenes, la minería de texto, reconocimiento de voz, detección de intrusos y bioinformática.

El funcionamiento de SVM implica el mapeo de los datos a un espacio de características de altas dimensiones para permitir la categorización de los puntos de datos, inclusive cuando estos datos no pueden ser separados linealmente. En

este proceso, se identifican separadores entre las diferentes categorías y se realizan una transformación de los datos para que los separadores puedan extraerse como hiperplanos. Posteriormente, para predecir a qué grupo pertenece un nuevo registro se podrán utilizar las características de los nuevos datos.

Supóngase la Figura 4, se presenta el conjunto de datos original, donde se tienen dos categorías diferentes representadas por puntos.

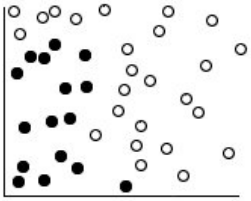


Figura 4. Conjunto de datos original

Como se observa en la Figura 5, para separar estas categorías se introduce una curva.

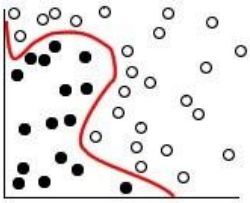


Figura 5. Datos con un separador

Posteriormente, luego de la transformación, mediante un hiperplano se define el límite entre las dos categorías, ilustrado en la Figura 6.

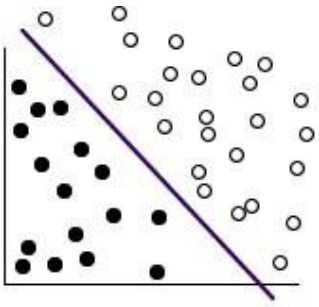


Figura 6. Datos transformados

La función matemática que se utilizada para la transformación de datos en SVM se conoce como función **kernel**. SVM generalmente soporta los siguientes tipos de kernel:

- Lineal
- Polinómico
- Función de base radial (RBF)
- Sigmoide

El uso de una función kernel lineal es aconsejable cuando los datos pueden ser separados linealmente de manera sencilla. En situaciones más complejas, se recomienda emplear alguna de las otras funciones mencionadas. Es necesario experimentar con las distintas funciones para encontrar el modelo más adecuado en cada situación, porque cada una utiliza algoritmos y configuraciones de parámetros distintas.

2.3. Toma de decisiones

Todas las personas tomamos decisiones todos los días de nuestra vida, en donde debemos escoger entre dos o más alternativas. Todas las decisiones siguen un proceso en común el cual puede ser descrito mediante pasos que se aplican a todas las circunstancias en las que toman decisiones, sean estas simples o complejas. (Robbins, 1987)

Existen diferentes procesos para llegar a una decisión, dependiendo de su importancia. Paul Moody (1983), describe este proceso como un circuito cerrado,

donde inicialmente halla un problema o discrepancia entre un estado deseado y la condición real. A continuación, se analizan las alternativas y consecuencias; una vez identificadas, cada una se evalúa de manera crítica en función de sus ventajas y desventajas. Para finalmente seleccionar e implementar una de las alternativas, con ello teniendo un curso de acción que dé resultados esperados, y proporcione una retroalimentación al proceso que ponga a prueba la validez y efectividad de la decisión.

Darío Parra (1998), expone que las decisiones se toman siguiendo las mismas etapas y están influenciadas por los mismos elementos que la acción. Donde una decisión comienza en la mente humana y finaliza en la ejecución de la acción correspondiente, pasando siempre por las etapas de deliberación y ejecución, es decir, un proceso cíclico influenciado por los conocimientos, expectativas y experiencias previas.

Con base en estas descripciones del proceso para llegar a una decisión se pueden identificar características importantes que influyen en la toma de decisiones las cuales son:

- **Información** la mayor parte de los casos se deben tomar decisiones con base en los datos disponibles y, aunque, tener una gran cantidad de información facilita el proceso de toma de decisión se deben tener presente el costo y beneficio de recolectar información.
- **Conocimientos,** ciertas decisiones necesitan de conocimientos específicos en caso de no tenerlos es necesarios buscar consejo en quienes estén informados.

- **Experiencia** cuando se cuenta con una mayor experiencia se podrán tomar decisiones instantáneas, gracias a los recuerdos de problemas y situaciones similares.
- **Análisis** tener un buen desarrollo de las capacidades analíticas es un factor importante en la toma de decisiones, sin embargo, cuando por medios de métodos analíticos no es posible solucionar el problema se puede y debe recurrir a la intuición.

Como se ha expuesto, la toma de decisiones es un proceso esencial para lograr el éxito de un proyecto y permite establecer dirección y rumbo a largo plazo. Así como es importante conocer los procedimientos para la tomar de decisiones y como llevar a cabo este proceso de manera efectiva y oportuna, para tener una mejor comprensión del papel de la información también es necesario profundizar en los temas relacionados con el uso, tratamiento de la información y las condiciones que favorecen estos aspectos.

2.4. Toma de decisiones en el marco de las Ciencias Computacionales

La toma de decisiones juega un papel fundamental en una amplia variedad de campos y disciplinas, su importancia en las Ciencias Computacionales es aún más relevante debido a la complejidad y la velocidad de cambio.

Una técnica para predecir la pertenencia de un objeto a un grupo, que es la variable dependiente, basándose en un conjunto de variables independientes llamadas variables predictoras, conocido como análisis discriminante introducido por Fischer en 1936. El objetivo de esta técnica es entender las discrepancias entre

los grupos y estimar la probabilidad de que un objeto pertenezca a una clase específica o grupo en función de los valores que toman las variables predictoras.

El análisis discriminante crea un modelo predictivo que incluye una función discriminante construida a partir de combinaciones lineales de variables predictivas ofreciendo la mejor separación posible dentro de los grupos. Esta función utiliza una muestra de casos donde se está al tal la pertenencia al grupo; posteriormente, a los nuevos casos donde se tienen mediciones de las variables predictoras se les puede aplicar esta función, ya que la pertenencia al grupo es desconocida (IBM, 2021).

Las variables de agrupación deben tener varias categorías y limitadas, codificadas con números enteros; y las variables independientes que son nominales deben recodificarse en variables auxiliares o de oposición. De igual manera, los casos deben ser independientes y se debe tener una distribución normal multivariada de las variables predictoras. Además, se considera que las matrices de varianzas-covarianzas dentro de cada grupo son las mismas. Se presupone que ningún caso pertenece a más de un grupo, es decir que es mutuamente exclusivo y que todos los casos pertenecen a uno.

CAPITULO 3

Modelado de Procesos

Con el propósito de comprender el proceso de asignación de áreas de conocimiento y contar con información precisa para el análisis, es fundamental identificar las actividades involucradas que permita representar cada fase de manera clara y estructurada. En este capítulo, se identificarán y detallaran las actividades del proceso, asimismo, se establecerá la variable objetivo y se analizará su comportamiento.

3.1. Análisis de los procesos de negocio

"*Solo quienes conocen sus procesos de principio a fin pueden optimizarlos, adaptarlos y alcanzar los objetivos con mayor rapidez y eficacia*" (GBTEC Software AG, 2024).

En la actualidad existe un entorno de alta competitividad entre las empresas, donde se les exigen que se adapten de manera eficaz a los cambios y finalicen las tares de forma eficiente, asegurado la mejora continua de sus procesos y la satisfacción del cliente interno y externo. Para lograr estos objetivos se recurren a modelos basados en los conceptos de gestión a través de sus procesos.

Hitpass (2017), define un proceso como un conjunto de actividades que se hacen bajo determinadas reglas con un fin, impulsadas por eventos. Con base en esta descripción se pueden identificar los elementos principales que describen un proceso:

- Los eventos son los que inician un proceso, ya que, algo tiene que ocurrir para que el proceso reaccione ante el suceso.
- Las actividades son acciones sobre un objeto, véase como el proceso de transformación. Además, se encuentran encadenas a una secuencia lógica que determinan las condiciones del negocio.

Hammer y Champy (1993), introducen el concepto de proceso de negocio como un conjunto de actividades que cuenta un uno o varios inputs que crean un output de valor para el cliente. De acuerdo con la definición de proceso que provee Hitpass, podemos definir que un proceso de negocio es un conjunto de actividades impulsadas por eventos y al ejecutarlas en secuencia, crean valor para el cliente (interno o externo).

En toda organización existen una gran cantidad de procesos de negocio donde se busca tener el mayor control y desempeño de estos, para obtener conocimiento en tiempo real de la carga de trabajo de cada usuario y reconocer aquellos procesos que se encuentran estancados, permitiendo detectar problemas antes que impacten en los resultados. Para observar, medir, controlar y analizar el desempeño de los procesos será necesario la introducción de la gestión por procesos, también conocidos como Business Process Management (BPM).

De acuerdo con Carraher (2013), la gestión por procesos (BPM) tiene la capacidad de mejorar la productividad y eficiencia, minimizar errores, reducir costos y proporcionar visibilidad sobre el cumplimiento de los objetivos y procesos. Además, se debe tener presente que en la actualidad el eje central de los procesos

es la tecnología de la información (IT), que garantiza contar con las aplicaciones y datos necesarios para funcionar.

Y en un mundo cambiante donde los modelos de negocio y las aplicaciones de IT existentes se deben modificar continuamente, es importante tener mecanismos que interconecten las estrategias y los procesos, manteniendo a la organización ágil y competitiva.

La definición que proporciona Association of BPM Professionals (ABPMP) sobre BPM es el de una disciplina integradora que abarca las capas de negocio y tecnología, comprometiéndolo como un todo integrado en gestión a través de los procesos.

Para aplicar BPM es necesario utilizar un ciclo de vida apegado a los principios de la Gestión de procesos de negocios, además la mayoría de los ciclos de vida son un conjunto de actividades, iterativo y por fases que incluyen:

- **Alineamiento con la estrategia y las metas:** Es la fase inicial donde se comprende las estrategias y objetivos de la organización que se diseñan para asegurar una propuesta de valor.
- **Diseño de cambios:** En esta fase se identifican los procesos principales e interfuncionales en el contexto de los objetivos y metas deseadas.
- **Desarrollar iniciativas:** Desarrollo de todos los planes para su implementación, donde se capacita sobre los procesos, los planes del

proyecto, gestión de cambios y obtención de beneficios se realizan en esta fase.

- **Implementar los cambios:** Se implementan todos los planes desarrollados de la fase desarrollar iniciativas. En esta fase se incluye un cronograma de implementación de proyecto estructurado en cada tarea y actividad con sus respectivas dependencias y predecesores.
- **Medir el éxito:** Se incluyen las mediciones con respecto a los beneficios proyectados del plan original y el monitoreo continuo de los procesos de negocio y de la tecnología.

Para modelar la colaboración entre los procesos, evitando una insuficiencia estructural al modelar la lógica se recomienda utilizar Business Process Management Notation (BPMN) que es la notación gráfica estandarizada para representar de forma secuencial las actividades a realizar en un proceso. Lo que asegura una comunicación eficiente, fluida y expresiva entre el personal del negocio y el personal técnico.

De acuerdo con Hitpass (2014), los elementos básicos de la notación BPMN son los objetos de flujo, que representan las actividades, eventos y condiciones (Gateways) del proceso, los objetos de conexión que como su nombre lo indica ayudar a conectar los objetos de flujo. Estos procesos solo pueden ser representados dentro de un lane que a su vez está dentro de un pool que representa a un participante. Además, para enriquecer de información a la descripción de un proceso podremos hacer uso de los artefactos, estos no tienen ninguna influencia en la lógica del proceso.

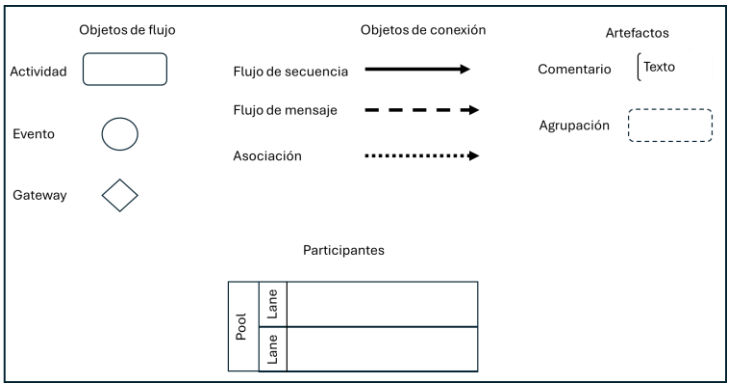


Figura 7. Elementos básicos de BPMN

3.2. Modelo de proceso de negocio para la clasificación de los objetos

Con lo anteriormente presentado se decidió implementar BPMN, ya que, permitirá identificar las actividades que existe dentro de cada proceso y plasma de manera clara la secuencia que debe seguir un objeto para lograr los resultados deseados de la organización.

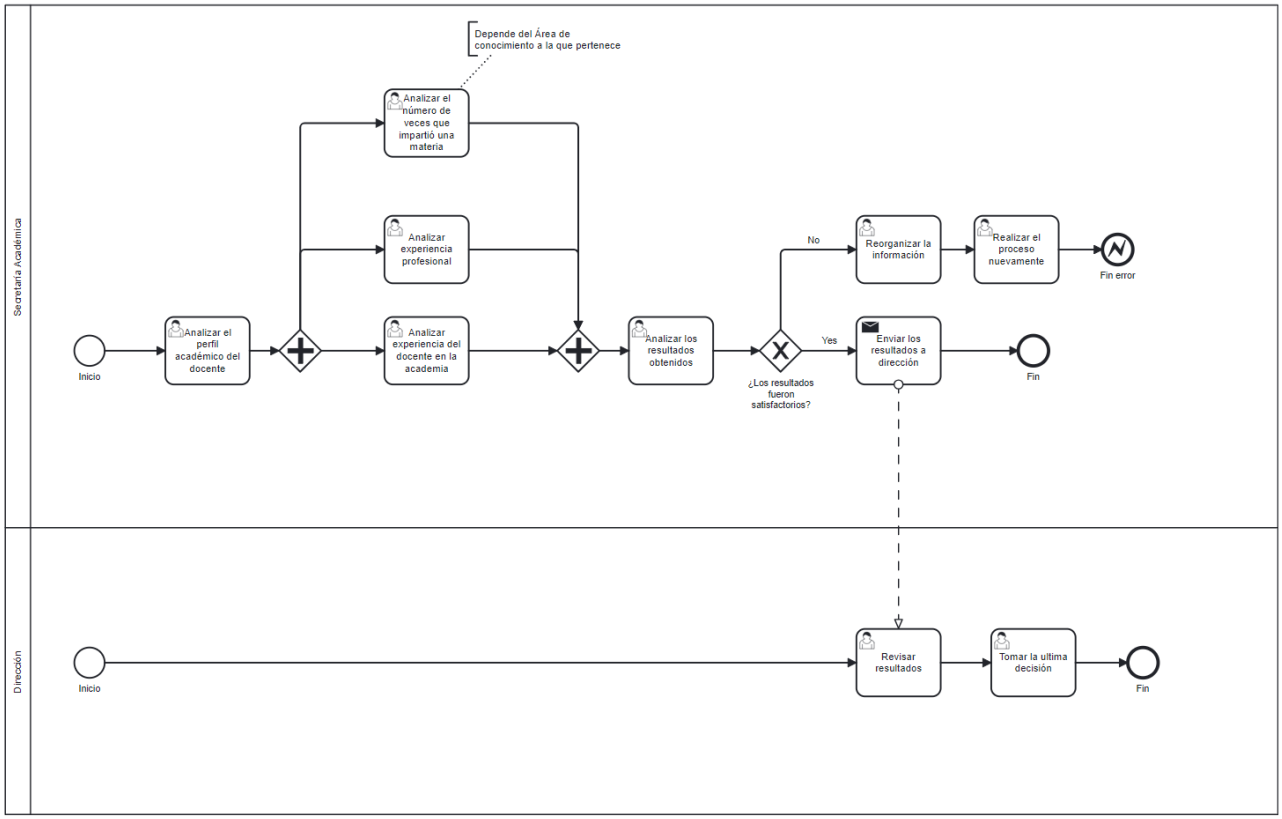


Figura 8. BPMN del proceso de negocio.

El diagrama de la Figura 8 constan de dos participantes, secretaria académica donde se realiza todo el proceso de asignación de área y dirección donde se aprueban los resultados obtenidos en la asignación de áreas por docente. La primera actividad que inicia el proceso es analizar el perfil académico del docente con el fin de asignarle un área de conocimiento adecuada a sus cualidades y dominios.

Para lograr una correcta clasificación se deben tener presente la experiencia del docente en la academia, su experiencia profesional y el número de veces que ha impartido una materia. Para representar este paralelismo (AND-Split) en BPMN se hace uso de un Gateway paralelo, además también debemos representar una sincronización de los flujos (AND-Join), en la Figura 9 se señala cada uno de estos.

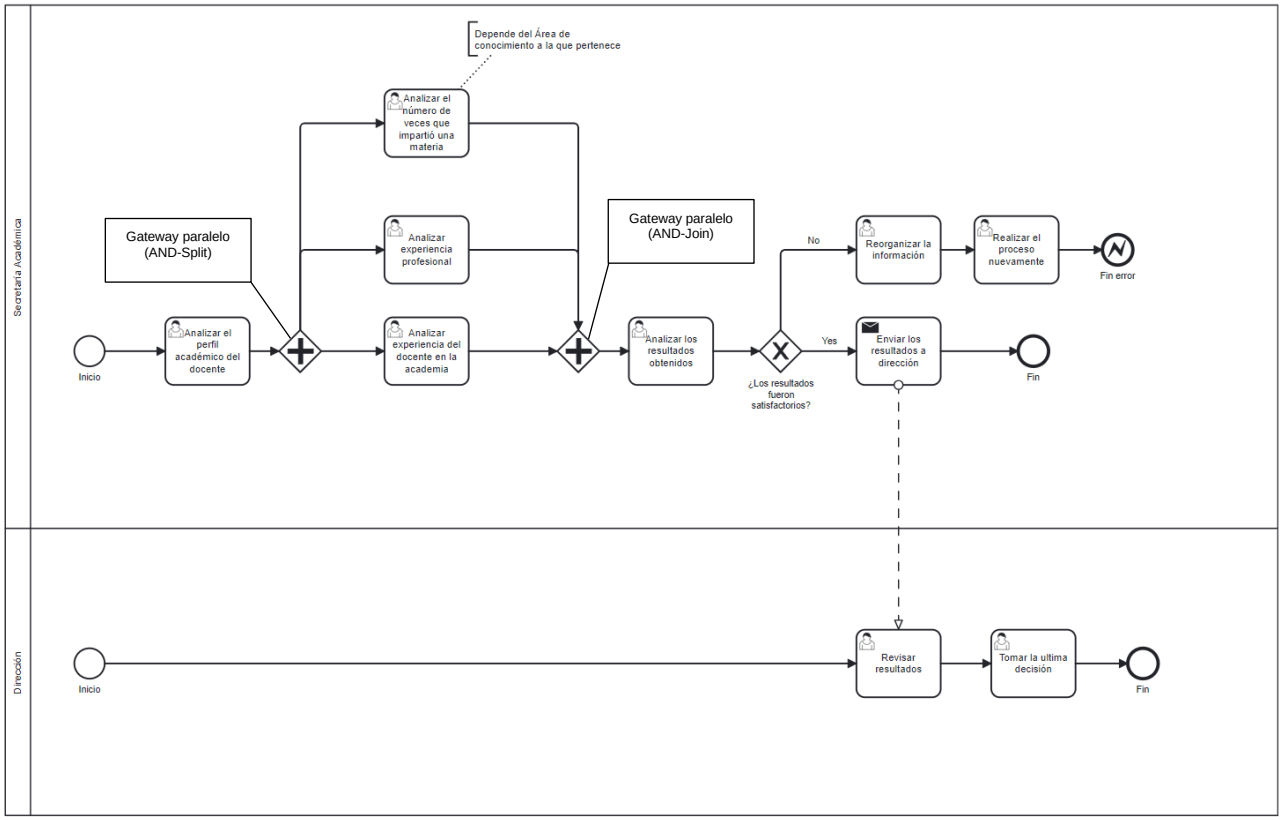


Figura 9. Gateway paralelo

El comportamiento que tendrá el token; que es una construcción teórica para visualizar y probar el comportamiento de los procesos, recorre la actividad de analizar el perfil académico del docente y llega al Gateway paralelo (AND-Split) donde se divide, el token se divide tantas veces como salidas tiene la compuerta, en este caso son tres. Cada uno de estos se dirigen a la actividad que les corresponde y una vez que esta actividad se termina se dirigen al AND-Join, donde la instancia debe esperar hasta que lleguen todos los tokens y los fusiona en uno sólo.

El flujo continúa con la actividad de analizar los resultados obtenidos de las actividades anteriores y posteriormente el token se dirige al XOR-Gateway, en donde según la decisión que se tome los resultados se envían a dirección donde estos serán aprobados o denegados, finalmente el token es consumido y desaparece, al mismo tiempo que la instancia muere. O en el caso que los resultados que se obtuvieron no son del todo acertados, se harán las actividades de reorganizar la información y se procese a realizar nuevamente el proceso, de igual manera el token desaparece y la instancia muere, pero con un error.

Con ayuda de BPMN se puede identificar que las actividades donde se toma más tiempo de realizar son en el análisis de todo el perfil académico de cada docente, pero de igual manera, se detectó que el analizar el número de veces que un docente impartió una materia puede optimizarse mediante un algoritmo de clasificación ayudando a una asignación más adecuada y eficiente.

CAPÍTULO 4

Construcción del modelo y predicciones

De acuerdo con lo planteado en el capítulo anterior la construcción de un modelo de clasificación utilizando aprendizaje supervisado será una importante herramienta importante para que la asignación de áreas correspondientes sea más eficiente. Ya que, al existir n posibles resultados (outputs) este se considera como una clasificación de objetos y es de aprendizaje supervisado porque en función de las variables dependientes se conoce el comportamiento de la variable objetivo.

Aunque existen un gran número de software que permite la construcción de modelos de minería de datos estos no permiten saber el funcionamiento interno. Por ello, una de las ventajas de trabajar con lenguajes de programación es la certeza de saber cómo funciona y que hace el algoritmo que se implementa, debido al importante desarrollo en el área de ciencias de datos que ha tenido Python se optó por utilizarlo para la construcción y evaluación del modelo.

Para entrenar modelos de machine learning es necesario tener cantidades de datos compatibles y precisos para lograr los objetivos establecidos en la organización. Por ello, es necesario recurrir a la integración de datos que consiste en combinar datos de múltiples orígenes brindando una visión unificada para una mejor toma de decisiones, una comprensión coherente de las operaciones y un mejor aprovechamiento de los datos, asimismo, comprender cómo funciona la

integración de datos es un aparte fundamental para lograr los mejores beneficios que nos aporta.

Existe una gran variedad de estrategias, la más común y utilizada es la consolidación de datos, que consiste en extraer, limpiar y almacenar datos en un entorno de almacenamiento de datos analítico, esta tiene dos tipos principales de herramientas, ETL que significa “extracción, transformación y carga” y ELT “extracción, carga y transformación”. Debido a que ETL realiza las transformaciones antes de cargar los datos en un entorno de almacenamiento, se asegura que los datos que ingresan al sistema ya están limpios y estructurados de acuerdo con los objetivos establecidos, gracias a esto se reduce la complejidad de las operaciones que deben realizarse dentro del entorno de almacenamiento, por ello, se decidió hacer uso de este tipo de consolidación de datos.

Asimismo, con ayuda de las herramientas que proporciona Excel y SQL Server se hará la limpieza y tratamiento para que los datos tengan los comportamientos adecuados para su utilización en los modelos de clasificación.

4.1. Exploración y visualización

Para entender mejor el comportamiento de los datos obtenidos es necesarios realizar como primer paso la visualización de la información con la que se cuenta. Los datos se recolectaron de una fuente externa que hacen referencia a la información necesaria para identificar el área de conocimiento de cada docente, esta contiene un total de 2126 registros y 8 columnas, los datos abarcan desde el año 2022 hasta 2023. A continuación, se muestran los datos recolectados con una descripción.

Atributo	Descripción
periodo	Indica los años correspondientes.
claves_programas	Es un identificador de cada una de las licenciaturas, se tienen las siguientes claves: LCC, ICC, ITI.
id_docente	Identificador único de cada docente.
materia	Almacena el nombre de la asignatura.
area_de_conocimiento	Indica el área de conocimiento a la que pertenece cada materia.
area_de_programa_educativo	Es el área que pertenece la materia de acuerdo con el mapa curricular de cada licenciatura.
area_de_formacion	Son los niveles de las licenciaturas, se tienen los siguientes: Básico, Formativo y Optativa.
clave	Es la variable objetivo, con rango: 8 a 13.

Se realizo una exploración minuciosa de los datos recolectados se identificó que algunos datos están vacíos, inconsistencias en las áreas de acuerdo con el mapa curricular de las licenciaturas, así como conjuntos de datos que hacen el análisis más complejo.

4.2. Selección de atributos

En esta etapa se seleccionaron los datos de acuerdo con las reglas del negocio, se asegura que los datos incluidos son relevantes y adecuados para las necesidades de la organización. Las reglas del negocio son esenciales para el éxito de cualquier proyecto ya que definen los criterios y restricciones que guiarán la selección, preparación y análisis de los datos, garantizando la calidad y pertenencia de los datos, además de optimizar la eficacia de los modelos de minería de datos

alineándolo con los objetivos específicos. Las principales características consideradas fueron:

- Claves del programa actuando como identificador único para cada una de las licenciaturas ofrecidas, que permiten una categorización clara y precisa de los programas académicos.
- Área de conocimiento fundamental para entender y analizar la distribución del contenido académico, así como identificar las áreas de especialización dentro de cada programa.
- Área de programa educativo mantienen la coherencia y alineación de las materias con los objetivos y estructuras de los programas académicos.
- Área de formación ayuda a estructurar y organizar el currículo académico, asegurando que la educación sea progresiva y equilibrada.

4.3. Tratamiento y limpieza

Como se mencionó anteriormente, se tienen datos faltantes e inconsistencias que podrían afectar a los resultados del modelo de minería de datos, es necesario realiza procedimientos para la preparación, limpieza y normalización de los datos del dataset, con la finalidad de asegurar su calidad y adecuación para el análisis posterior, los pasos específicos realizados se explican a continuación.

Como primer paso, se realizó la estandarización de los claves del programa. Inicialmente, estas claves contenían conjuntos de valores como {CCO,ICC}, {CCO,ITI}, {ICC,ITI}, {CCO, ICC, ITI}. Para simplificar el análisis se decidió convertir

estos conjuntos en varias filas diferentes, cada una con una clave única, es decir, solo contener tres tipos de claves que representan las licenciaturas ofertadas: CCO, ICC, ITI. Este proceso de estandarización provee una consistencia y precisión de los datos a lo largo del análisis, permitiendo una mejor categorización y manejo de la información.

Posteriormente se identificaron registros faltantes en las columnas “Área de conocimiento”, “Área de programa educativo” y “Área de formación”, estos registros faltantes fueron identificados con los valores de “0”, “PA” y “SIN ÁREA”. Para abordar este problema, se llevó a cabo un proceso de filtrado en el dataset con el fin de seleccionar las filas que contenían alguno de estos valores indicativos de datos vacíos donde finalmente se procedió a eliminar estas filas del dataset. Con ello se asegura la integridad y calidad de los datos, ya que la presencia de registros incompletos podría afectar negativamente el análisis.

Debido a la normalización de las claves de los programas, se realizaron ajustes significativos en las áreas de programa educativo. De acuerdo con los mapas curriculares de las licenciaturas ofertadas, se identificaron tres áreas comunes: “Área de Ciencias Básicas”, “Área de Tecnología” y “Área de Formación General Universitaria”. Además de estas áreas en común, cada licenciatura cuenta con un área adicional específica. Para Ingeniería en Tecnologías de la Información (ITI), se incluye el “Área de Modelado de Sistemas”, para Ingeniería en Ciencias de la Computación (ICC), se tiene el “Área de Ingeniería en Computación”; y para la Licenciatura en Ciencias de la Computación (CCO), se añade el “Área de Ciencias de la Computación”.

En este contexto, se realizaron los ajustes necesarios para garantizar la coherencia y precisión de las áreas asignadas a cada registro en el dataset. Para aquellos registros que contenían combinaciones de áreas como “Área de ingeniería en computación / Área de ciencias de la computación” o “Área de modelado de sistemas / Área de tecnología”, se procedió a asignar el área correspondiente de acuerdo con la clave de programa especificada en cada registro. De igual manera, se identificaron y corrigieron aquellos registros que presentaban un área errónea en relación con su clave de programa. Permitiendo una representación fiel y adecuada de la estructura establecida por los mapas académicos de las licenciaturas.

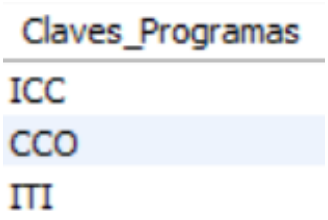
Por otro lado, el área de formación inicialmente contaba con conjuntos de valores como {Básico, Formativo}, {Formativo, Desit} y {Formativo, Optativa}. Se decidió unificar los conjuntos {Formativo, Desit} y {Formativo, Optativa} bajo el valor único de “Formativo”. Asimismo, se cambiaron los registros que contenían los valores “Optativa” y “Optativa Desit” a “Formativo”, garantizando una categorización uniforme.

Al analizar el conjunto de valores {Básico, Formativo}, se descubrió que algunas materias, específicamente “Graficación” y “Sistemas Operativos II”, se ofertaban en las tres licenciaturas, pero en diferentes semestres y niveles de formación. En el caso de “Graficación” se impartía en el nivel básico para las licenciaturas de Ingeniería en Ciencias de la Computación (ICC) e Ingeniería en Tecnologías de la Información (ITI), mientras que para la Licenciatura en Ciencias de la Computación (CCO), se ofertaba en el nivel formativo. Por otro lado, la materia “Sistema Operativos II” se categorizó como nivel formativo tanto pata ICC como para

CCO. Estas modificaciones aseguraron que cada materia esté correctamente clasificada según su nivel de formación apropiado para cada licenciatura.

Para entrenar un modelo de aprendizaje supervisado es esencial que los datos no sean categóricos. Por lo tanto, se realizó la transformación de los datos originales del dataset a valores numéricos. Este proceso implicó contar cada categoría específica y convertirla en un valor numérico correspondiente.

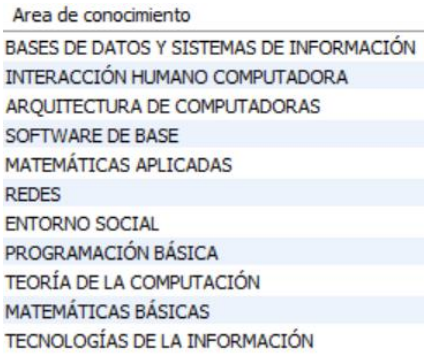
Con ayuda de SQL Server, los datos se importaron los datos con las debidas transformaciones. Para las claves de programas, se definieron tres tipos como se muestran en la Figura 10.



Claves_Programas
ICC
CCO
ITI

Figura 10. Claves del programa

Asimismo, se identificó los diferentes valores que contiene cada una de las áreas de conocimiento, área de programa educativo y área de formación. Los resultados de estas identificaciones se presentan en las Figuras 11, 12 y 13.



Area de conocimiento
BASES DE DATOS Y SISTEMAS DE INFORMACIÓN
INTERACCIÓN HUMANO COMPUTADORA
ARQUITECTURA DE COMPUTADORAS
SOFTWARE DE BASE
MATEMÁTICAS APLICADAS
REDES
ENTORNO SOCIAL
PROGRAMACIÓN BÁSICA
TEORÍA DE LA COMPUTACIÓN
MATEMÁTICAS BÁSICAS
TECNOLOGÍAS DE LA INFORMACIÓN

Figura 11. Áreas de conocimiento

Area de programa educativo
AREA DE TECNOLOGIA
AREA DE INGENIERIA EN COMPUTACION
AREA DE FORMACION GENERAL UNIVERSITARIA
AREA DE CIENCIAS BASICAS
AREA DE CIENCIAS DE LA COMPUTACION
AREA DE MODELADO DE SISTEMAS

Figura 12. Áreas de programa educativo

Area de formacion
FORMATIVO
BASICO

Figura 13. Áreas de formación

A continuación, se realizó el conteo de los datos utilizando el siguiente query SQL, ilustrado en la Figura 14. Este query permitió transformar las categorías en datos numéricos.

```
SELECT id_docente, clave_cua,
    SUM(CASE WHEN claves_programas = 'ICC' THEN 1 ELSE 0 END) AS 'ICC',
    SUM(CASE WHEN claves_programas = 'CCO' THEN 1 ELSE 0 END) AS 'CCO',
    SUM(CASE WHEN claves_programas = 'ITI' THEN 1 ELSE 0 END) AS 'ITI',
    SUM(CASE WHEN area_de_formacion = 'BASICO' THEN 1 ELSE 0 END) AS 'BASICO',
    SUM(CASE WHEN area_de_formacion = 'FORMATIVO' THEN 1 ELSE 0 END) AS 'FORMATIVO',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE CIENCIAS BASICAS' THEN 1 ELSE 0 END) AS 'AREA DE CIENCIAS BASICAS',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE MODELADO DE SISTEMAS' THEN 1 ELSE 0 END) AS 'AREA DE MODELADO DE SISTEMAS',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE TECNOLOGIA' THEN 1 ELSE 0 END) AS 'AREA DE TECNOLOGIA',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE FORMACION GENERAL UNIVERSITARIA' THEN 1 ELSE 0 END) AS 'AREA DE FORMACION GENERAL UNIVERSITARIA',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE CIENCIAS DE LA COMPUTACION ' THEN 1 ELSE 0 END) AS 'AREA DE CIENCIAS DE LA COMPUTACION ',
    SUM(CASE WHEN area_de_programa_educativo = 'AREA DE INGENIERIA EN COMPUTACION' THEN 1 ELSE 0 END) AS 'AREA DE INGENIERIA EN COMPUTACION',
    SUM(CASE WHEN area_de_conocimiento = 'ENTORNO SOCIAL' THEN 1 ELSE 0 END) AS 'ENTORNO SOCIAL',
    SUM(CASE WHEN area_de_conocimiento = 'TEORIA DE LA COMPUTACION' THEN 1 ELSE 0 END) AS 'TEORIA DE LA COMPUTACION',
    SUM(CASE WHEN area_de_conocimiento = 'INTERACCION HUMANO COMPUTADORA' THEN 1 ELSE 0 END) AS 'INTERACCION HUMANO COMPUTADORA',
    SUM(CASE WHEN area_de_conocimiento = 'ARQUITECTURA DE COMPUTADORAS' THEN 1 ELSE 0 END) AS 'ARQUITECTURA DE COMPUTADORAS',
    SUM(CASE WHEN area_de_conocimiento = 'TECNOLOGIAS DE LA INFORMACION' THEN 1 ELSE 0 END) AS 'TECNOLOGIAS DE LA INFORMACION',
    SUM(CASE WHEN area_de_conocimiento = 'MATEMATICAS BASICAS' THEN 1 ELSE 0 END) AS 'MATEMATICAS BASICAS',
    SUM(CASE WHEN area_de_conocimiento = 'REDES' THEN 1 ELSE 0 END) AS 'REDES',
    SUM(CASE WHEN area_de_conocimiento = 'MATEMATICAS APLICADAS' THEN 1 ELSE 0 END) AS 'MATEMATICAS APLICADAS',
    SUM(CASE WHEN area_de_conocimiento = 'BASES DE DATOS Y SISTEMAS DE INFORMACION' THEN 1 ELSE 0 END) AS 'BASES DE DATOS Y SISTEMAS DE INFORMACION',
    SUM(CASE WHEN area_de_conocimiento = 'SOFTWARE DE BASE' THEN 1 ELSE 0 END) AS 'SOFTWARE DE BASE',
    SUM(CASE WHEN area_de_conocimiento = 'PROGRAMACION BASICA' THEN 1 ELSE 0 END) AS 'PROGRAMACION BASICA'
    INTO base_de_datos_final
    FROM base_datos_2022_2023
    GROUP BY id_docente, clave_cua
    ORDER BY clave_cua;
```

Figura 14. Query de consulta

Los resultados de la trasformación se almacenaron en una nueva tabla llamada “base_de_datos_final”. Esto facilita el acceso y uso de los datos transformados para el entrenamiento del modelo de aprendizaje supervisado.

4.4. Entrenamiento del modelo y obtención de predicciones

Finalmente combinando todos los elementos anteriormente mencionados que permita determinar cuál es el mejor subconjunto de variables, los hiperparámetros óptimos de cada modelo y determinar qué modelo obtiene el mejor accuracy. Inicialmente, se consideraron tres modelos de aprendizaje supervisado: K-Nearest Neighbors (KNN), árboles de clasificación y Support Vector Machine (SVM). Cada uno fue seleccionado por sus características y potencial para manejar el tipo de datos disponibles en el dataset.

Cada uno de estos modelos fue entrenado y evaluado utilizando métricas predictivas: precisión, recall, F1-score y la matriz de confusión para garantizar la robustez de los resultados. Estas son fundamentales en la evaluación de modelos de clasificación ya que proporcionan información valiosa y complementaria sobre el desempeño del modelo. Para entrenar cualquier modelo de aprendizaje supervisado es necesario dividir el conjunto de datos en dos partes: un conjunto de entrenamiento utilizado para entrenar el modelo y un conjunto para pruebas que es usado como prueba para el modelo después de ser entrenado, este nos permite calcular el rendimiento y la evaluar el modelo.

De acuerdo con Microsoft (2023) la precisión indica el número de verdaderos positivos o negativos, indicando la frecuencia que se realizan predicciones correctas. La fórmula para calcular la precisión es la siguiente:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

La coincidencia; también conocido como recall, revela cuántas de las clases están correctamente etiquetadas. Al obtener un puntaje bajo de coincidencia indica que el modelo predice menos positivos verdaderos. La fórmula de puntación de coincidencia es:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Y el F1-score es el promedio armónico de la precisión y la coincidencia, además ayuda a encontrar un equilibrio entre estas dos. Esta métrica puede tener dos valores 1 que simboliza un ideal imposible, y 0 indicando que el algoritmo fallo todo el tiempo. El F1-score se calcula de la siguiente forma:

$$F1 = \frac{Recall}{Precision + Recall}$$

La matriz de confusión muestra las predicciones del modelo frente a las verdaderas etiquetas, detallando los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

A continuación, se presentan los resultados para los tres modelos que se obtuvieron después del entrenamiento y correspondiente prueba, de igual manera, se presenta la interpretación de las gráficas de las métricas predictivas y la matriz de confusión.

Para el modelo de KNN la combinación optima de hiperparámetros con un accuracy de 0.64: **{'metric': 'euclidean', 'n_neighbors':7, 'weights': 'distance'}**, el modelo obtuvo un buen rendimiento para las clases 8, 9 y 10, con buenos niveles de precisión y recall. La clase 11 muestra un rendimiento adecuado, sin embargo,

las clases 12 y 13 tienen un rendimiento significativamente bajo, en la Figuras 15 se muestras las métricas predictivas y en la Figura 16 se adjunta un reporte de los resultados. La Figura 17 muestra la matriz de confusión.

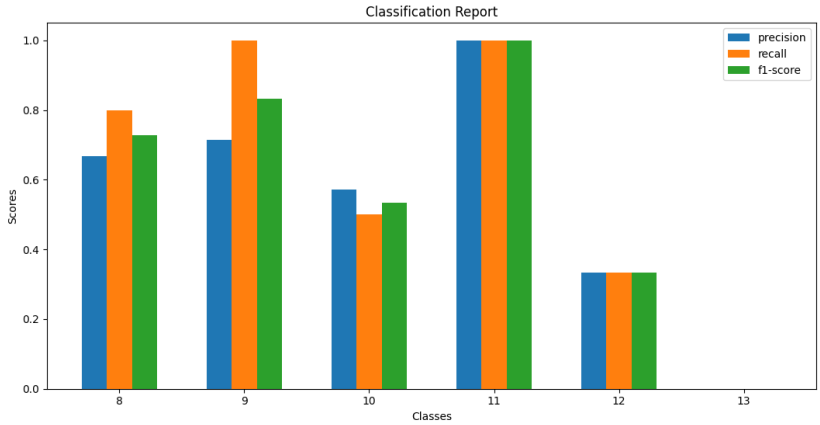


Figura 15. Métricas predictivas del modelo KNN.

```
KNN Best Parameters: {'metric': 'euclidean', 'n_neighbors': 7, 'weights': 'distance'}
Accuracy: 0.64
Classification Report:
              precision    recall  f1-score   support

           8       0.67      0.80      0.73         5
           9       0.71      1.00      0.83         5
          10       0.57      0.50      0.53         8
          11       1.00      1.00      1.00         2
          12       0.33      0.33      0.33         3
          13       0.00      0.00      0.00         2

    accuracy                           0.64        25
   macro avg       0.55      0.61      0.57        25
weighted avg       0.58      0.64      0.60        25
```

Figura 16. Reporte de las métricas predictivas del modelo KNN.

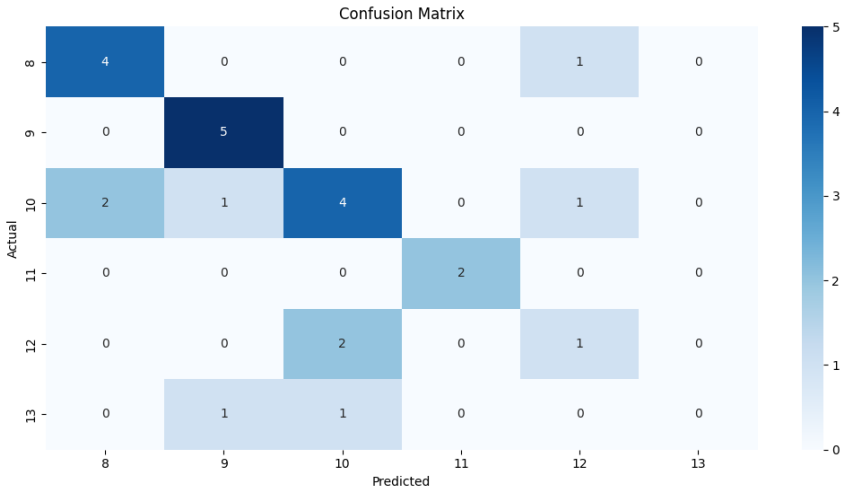


Figura 17. Matriz de confusión del modelo KNN.

Como segundo modelo entrenado fue árboles de clasificación dando un accurancy de 0.40, la combinación óptima para el modelo fue **{‘criterion’: ‘gini’, ‘max_depth’: 10, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2}**, el modelo presenta un rendimiento variable dependiendo de la clase. Las clases con buenos niveles de precisión y recall fueron la clase 8 y 9, asimismo la clase 11 muestra un buen rendimiento general. Por el contrario, las demás clases tiene un desempeño considerablemente bajo. A continuación, las Figuras 18, 19 y 20 muestran los resultados obtenidos.

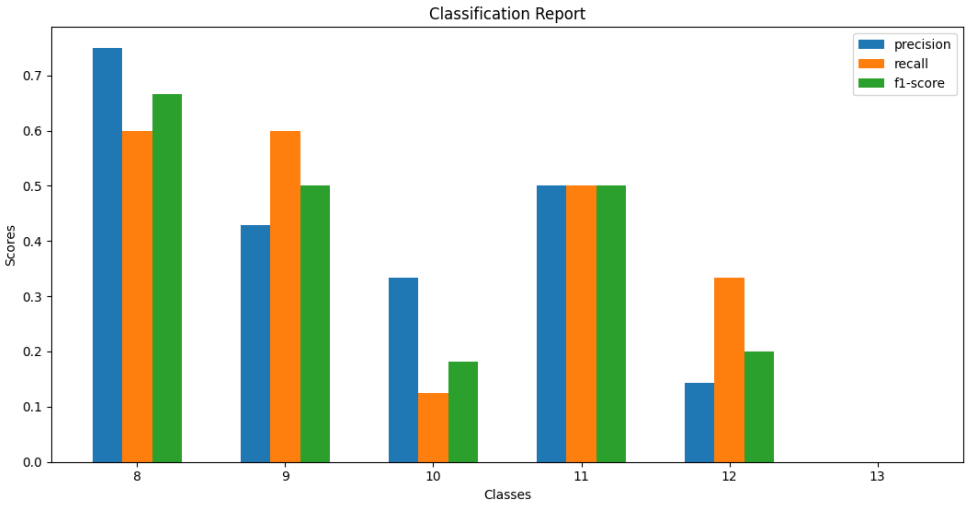


Figura 18. Métricas predictivas del modelo arboles de clasificación.

```
Decision Tree Best Parameters: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5}
Accuracy: 0.4
Classification Report:
              precision    recall  f1-score   support

           8       0.60      0.60      0.60         5
           9       0.50      0.60      0.55         5
          10       0.25      0.12      0.17         8
          11       1.00      0.50      0.67         2
          12       0.25      0.67      0.36         3
          13       0.00      0.00      0.00         2

    accuracy                           0.40        25
   macro avg       0.43      0.42      0.39        25
weighted avg       0.41      0.40      0.38        25
```

Figura 19. Reporte de las métricas predictivas del modelo arboles de clasificación.

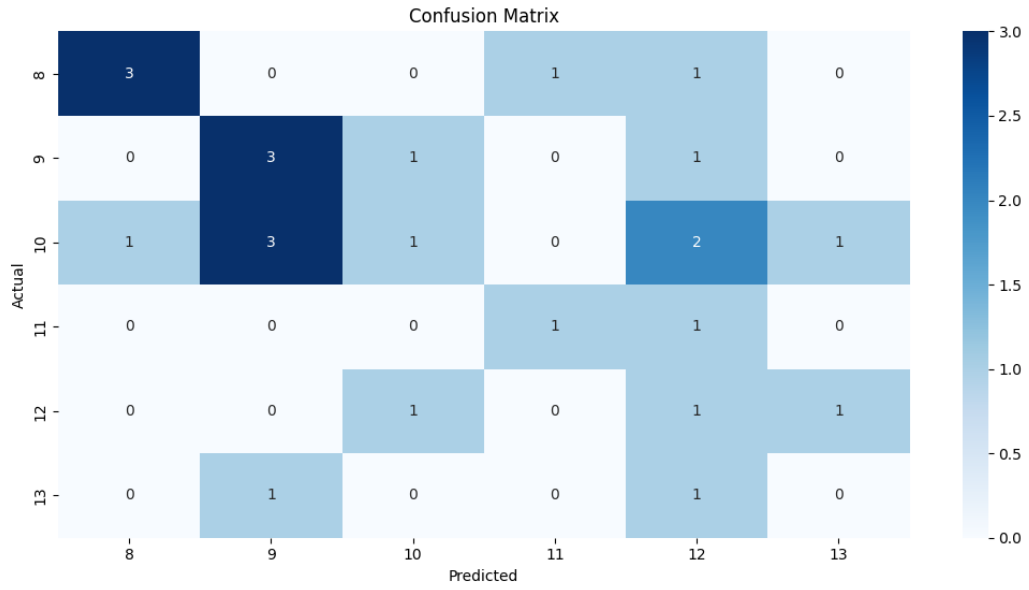


Figura 20. Matriz de confusión del modelo arboles de clasificación.

Como último modelo a entrenar fue Máquina de Vectores de soporte (SVM) la mejor combinación para este modelo fue **{‘C’: 1, ‘degree’: 2, ‘kernel’: ‘linear’}** dando un accurancy de 0.60, los resultados para las clases 8 y 11 muestran un excelente rendimiento, con altos niveles de precisión y recall, de igual manera, las clases 9 y 12 tienen un rendimiento aceptable. La clase 10 muestra un rendimiento moderado con varios errores de clasificación.

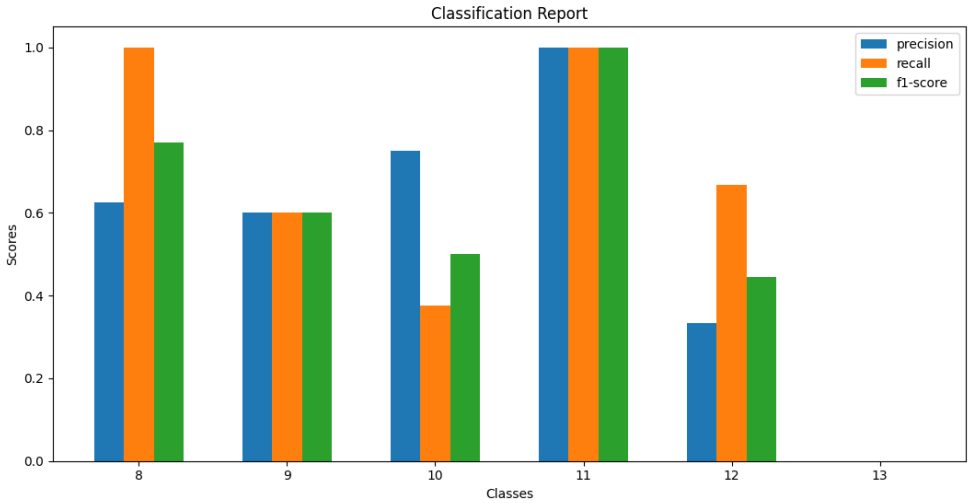


Figura 21. Métricas predictivas del modelo SVM.

```
SVM Best Parameters: {'C': 1, 'degree': 2, 'kernel': 'linear'}
Accuracy: 0.6
Classification Report:
              precision    recall  f1-score   support

           8       0.62      1.00      0.77         5
           9       0.60      0.60      0.60         5
          10       0.75      0.38      0.50         8
          11       1.00      1.00      1.00         2
          12       0.33      0.67      0.44         3
          13       0.00      0.00      0.00         2

    accuracy                           0.60        25
   macro avg       0.55      0.61      0.55        25
weighted avg       0.60      0.60      0.57        25
```

Figura 22. Reporte de métricas predictivas del modelo SVM.

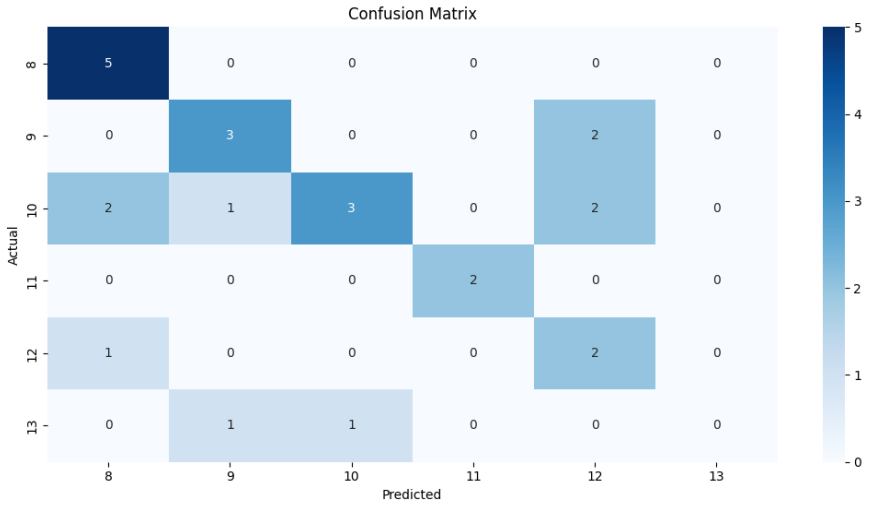


Figura 23. Matriz de confusión del modelo SVM.

A pesar de que los modelos muestran un rendimiento aceptable para la clasificación de los datos, esta puede mejorar considerablemente. Haciendo uso de Stacking Classifier podremos combinar las fortalezas de varios modelos base para mejorar el rendimiento de las predicciones y reducir la variabilidad. Esto permitirá abordar mejor las clases difíciles y mejorar las métricas de precisión y recall en general, proporcionando un modelo más robusto y preciso para la tarea de clasificación.

Una vez entrenado el modelo de Stacking Classifier los modelos bases que mejor rendimiento mostraron fueron SVM **{kernel='linear', C=4.0}** y **{kernel='poly', degree=3, C=3.0}** y como meta_estimator utilizado fue RandomForestClassifier. Obtuvo un accurancy de 0.76, en las clases 8, 9, 11 y 12 muestra una alta precisión y recall, lo que indica que el modelo es muy efectivo para clasificar correctamente estas clases. Además, que la clase 10 tiene un rendimiento aceptable, con precisión y recall moderados.

El modelo Stacking Classifier mostro ser altamente efectivos en varias clases, además de tener puntos fuertes en clasificarlas. Las Figuras 24, 25 y 26 muestran los resultados del modelo.

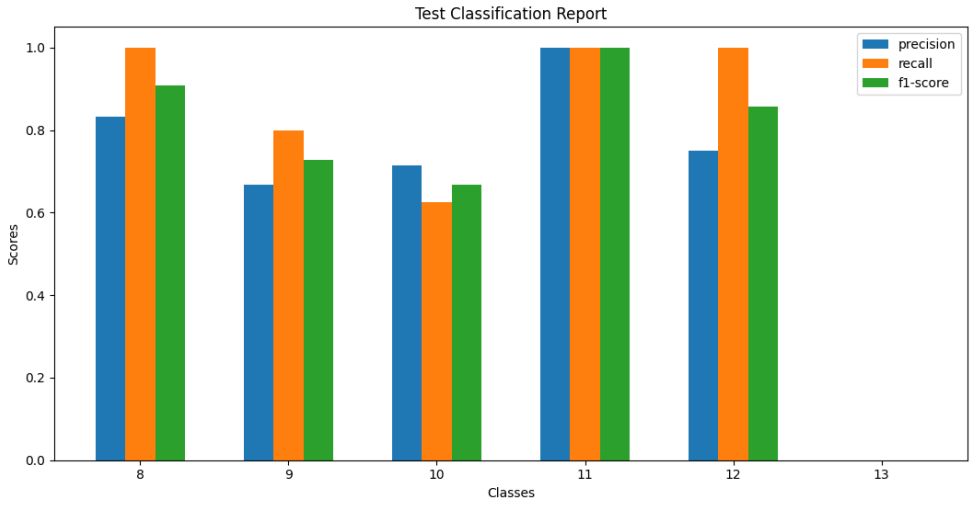


Figura 24. Métricas predictivas del modelo Stacking Classifier.

	precision	recall	f1-score	support
8	0.83	1.00	0.91	5
9	0.67	0.80	0.73	5
10	0.71	0.62	0.67	8
11	1.00	1.00	1.00	2
12	0.75	1.00	0.86	3
13	0.00	0.00	0.00	2
accuracy			0.76	25
macro avg	0.66	0.74	0.69	25
weighted avg	0.70	0.76	0.72	25

```
Test Confusion Matrix:
[[5 0 0 0 0 0]
 [0 4 1 0 0 0]
 [1 1 5 0 1 0]
 [0 0 0 2 0 0]
 [0 0 0 0 3 0]
 [0 1 1 0 0 0]]
Precisión del modelo Stacking: 76.00%
```

Figura 25. Reporte de las métricas predictivas del modelo Stacking Classifier.

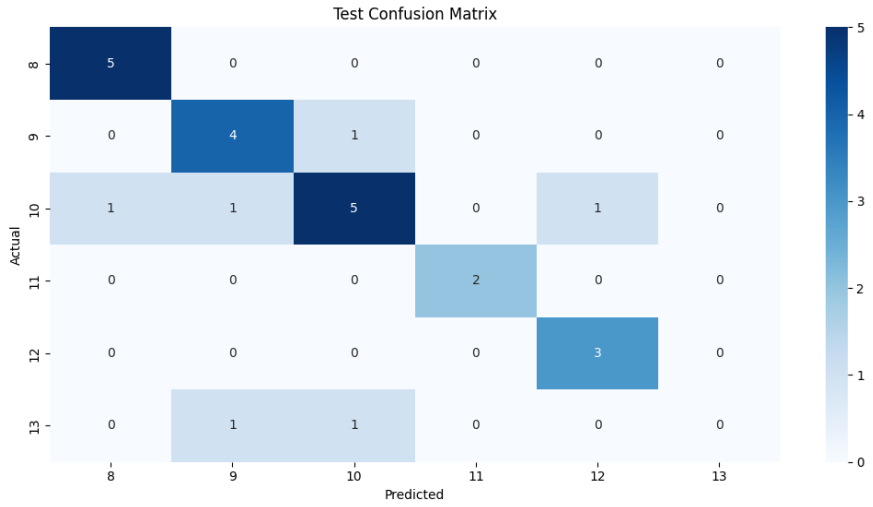


Figura 26. Matriz de confusión del modelo de Stacking Classifier.

En este estudio se evaluaron cuatro modelos de clasificación: KNN tiene un rendimiento aceptable en ciertas clases y aunque pude ser útil en algunos escenarios, su desempeño inconsistente limita su eficacia como modelo principal. Los árboles de clasificación ofrecen una buena interpretabilidad y manejan bien características mixtas, sin embargo, debido a su rendimiento inconsistente y confusión entre clases lo hace menos efectivo. SVM al ser un modelo robusto y consistente fue especialmente eficaz en manejar datos de alta dimensionalidad, su capacidad de generalización y precisión lo hacen una opción fuerte para el modelo de datos.

El modelo de Stacking Classifier al combinar las fortalezas de los modelos individuales logra un rendimiento superior y más consistente. Muestra una mejora significante en el rendimiento global y es particularmente eficaz en clases con más datos. No obstante, es necesario abordar las limitaciones que tuvieron todos los modelos en las clases minoritarias, como la clase 13, es necesario recolectar más datos y con ello, optimizar los modelos.

Capítulo 5

Resultados

Se presentan los resultados de la investigación, el impacto socioeconómico y las conclusiones.

5.1. Resultados

Un modelo de clasificación que determina, con una eficiencia mayor al 70%, la pertenencia de los objetos a una clase; esto, mediante técnicas de Machine Learning, siendo los objetos de estudio personas que tienen un perfil académico de Ciencias Computacionales y las clases un conjunto de Áreas de Conocimiento definidas por el usuario final.

La implementación del modelo podrá proporcionar información valiosa para la toma de decisiones, ayudando a los responsables a tener decisiones informadas sobre como asignar los recursos. Además de mejorar el rendimiento de los sistemas al garantizar una asignación precisa de objetos a clase. Al elegir un algoritmo de clasificación adecuado se pueden reducir los errores de asignación, minimizando los riesgos asociados con una asignación errónea o inadecuada.

5.2. Impacto Socioeconómico

Con la implementación del modelo de clasificación se podrán tener asignaciones más rápidas y precisas, beneficiando a las instituciones educativas de nivel superior ya que permite la clasificación de los académicos en las áreas donde mejor se ajusten sus conocimientos y habilidades.

Conclusiones

En este trabajo se ha logrado implementar un modelo de clasificación basado en técnicas de Machine Learning, que permite mejorar de manera significativa la asignación de áreas de conocimiento a los docentes de Ciencias Computacionales. El modelo final, superó las expectativas iniciales al alcanzar una precisión del 76%, lo que evidencia su capacidad para categorizar de manera precisa a los docentes dentro de áreas específicas según su perfil académico.

El proceso de construcción del modelo involucró una serie de pasos clave, desde la limpieza y normalización de datos hasta la evaluación exhaustiva de diversos algoritmos, lo que permitió asegurar que el modelo final fuera robusto y adecuado. La implementación de este modelo proporciona beneficios significativos, como la reducción de errores en la asignación, la optimización del tiempo y recursos, y un impacto directo en la calidad educativa de nivel superior.

Es importante destacar que este modelo no solo facilita la asignación de áreas de conocimiento, sino que también respalda la toma de decisiones informadas por parte de los responsables académicos, garantizando que los docentes sean asignados a áreas en las que mejor se ajusten sus habilidades y experiencia

Como trabajo futuro, se propone recolectar y analizar más datos para mejorar el rendimiento del modelo, especialmente en aquellas clases donde se observó un rendimiento bajo debido a la limitada disponibilidad de datos. La combinación de técnicas y la mejora continua de los modelos serán claves para seguir avanzando en este campo, asegurando que la tecnología siga siendo un apoyo crucial en la gestión educativa y en la mejora de los procesos institucionales.

Referencias

1. Amazon (2024). ¿Cuál es la diferencia entre machine learning supervisado y el no supervisado? Recuperado de: https://aws.amazon.com/es/compare/the-difference-between-machine-learning-supervised-and-unsupervised/
2. Caro-Castro, Carmen (2010). Las clasificaciones bibliográficas: de los estantes a la web. Tábula, 13, p.11-23
3. Carraher, K. (2013). Why / Por qué BPM (Business Process Manage-ment). Recuperado de: http://issuu.com/bpmteca/docs/ebook_why_porque_bpm_75c3d2cbb7a072
4. Eckert, Karina B, & Suénaga, Roberto. (2015). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. Formación universitaria, 8(5), 03-12. https://dx.doi.org/10.4067/S0718-50062015000500002
5. García R. (2015). Estudio comparativo entre las metodologías ágiles y las metodologías tradicionales para la gestión de proyectos de software. Universidad de Oviedo. https://digibuo.uniovi.es/dspace/bitstream/handle/10651/32457/TFMMlJGarciaRodriguezRUO.pdf;jsessionid=E7D9B15F52D0A0E1EDE64312C6A1F330?sequence=6
6. GBTEC Software AG (2024). ¿Qué es BPM? Definición y aplicaciones. Recuperado de: https://www.gbtec.com/es/recursos/bpm/

7. Gonzales Casanova, P. (1996). “El colonialismo global y la democracia”. Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades-Universidad Nacional Autónoma de México, Barcelona.
8. Hammer, M. y Champy, J. (1993). Reingeniería. Bogotá: Norma.
9. Haro S., (2018). Métodos de clasificación en minería de datos meteorológicos. Escuela superior politécnica de chimborazo, Riobamba. Ecuador, 20(2), 107-113. ISSN 2477-9105. Recuperado de: http://ceaa.espoch.edu.ec:8080/revista.perfiles/faces/Articulos/Perfiles20Art13.pdf
10. Hitpass, B. (2017). BPM: Business Process Management: Fundamentos y Conceptos de Implementación. Createspace Independent Publishing Platform; Edición 4.
11. IBM, (2021). Discriminant Analysis. Recuperado de: https://www.ibm.com/docs/es/spss-statistics/beta?topic=features-discriminant-analysis
12. IBM, (2021). Funcionamiento de SVM. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works
13. IBM, (2023). ¿Qué es KNN? Recuperado de: https://www.ibm.com/mx-es/topics/knn#:~:text=El%20algoritmo%20de%20k%20vecinos,un%20punto%20de%20datos%20individual.
14. IBM, (2023). Árbol de clasificación. Recuperado de: https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-classification-tree

15. INEGI, (2005). Principios básicos de las clasificaciones estadísticas en el ámbito sociodemográfico – Histórica. Recuperado de: https://www.inegi.org.mx/contenidos/clasificadoresycatalogos/doc/principios_basicos_de_las_clasificaciones.pdf
16. Ingry-Nathaly Salamanca R, Edgar Junior C., (2021), Técnicas de aprendizaje automático aplicadas en los sistemas de predicción. Tecnol.Investig. Academia TIA, ISSN: 2344-8288, 8 (1), pp. 39-53. Bogotá-Colombia.
17. Medina, R.D. (2012). Algunas reflexiones sobre la clasificación de los organismos vivos. Historia Ciencias Saude-manguinhos, 19, 883-898.
18. Molina R., Honores T., Pedreira S. y Pardo L. (2021). Estado del arte: Metodologías de desarrollo de aplicaciones móviles. 3C Tecnología. Glosas de innovación aplicadas a la pyme. ISSN: 2254 – 4143. Ed. 38 Vol. 10 N.º 2 Junio - Septiembre 2021
19. Moody, Paul E. (1983). Decision making: methods for better decisions. New York. Mc. GrawHill.
20. Nuñez-Lira, L. A., Alfaro Bernedo, J. O., Aguado Lingan, A. M., y González Ponce de León, E. R. (2023). Toma de decisiones estratégicas en empresas: Innovación y competitividad. Revista Venezolana De Gerencia, 28(No. Especial 9), 628-641. Recuperado de: https://doi.org/10.52080/rvgluz.28.e9.39.
21. Parra, D. (1998). Los modelos de decisión y la práctica del empresario frente a la toma de decisiones: esquema teórico y estudio en la empresa colombiana. Tesis doctoral. Medellín.

22. Parra, F., (2019). Estadística y Machine Learning con R. Recuperado de: https://bookdown.org/content/2274/portada.html
23. Raschka, S., (2018). STAT 479: Machine Learning Lectures Notes. University of Wisconsin–Madison. Recuperado de: https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/02_knn_notes.pdf
24. Robbins, S. (1987). Administración teórica y práctica. Prentice-Hall Hisponoamerica S.A. México.
25. RODRÍGUEZ RODRíGUEZ, J. E., (2007). Software para clasificación/predicción de datos. Tecnura, 11(21),41-53. ISSN: 0123-921X. Recuperado de: https://www.redalyc.org/articulo.oa?id=257021008004
26. Rodríguez Y., Díaz A. (2009). Herramientas de minería de datos. Revista Cubana de Ciencias Informáticas, 3(3-4),73-80. ISSN: 1994-1536. Recuperado de: https://www.redalyc.org/articulo.oa?id=378343637009
27. Rodríguez Y., Pinto M. (2010). Evolución, particularidades y carácter informacional de la toma de decisiones organizacionales. Revista Cubana de Información en Ciencias de la Salud, 21(1),57-77. Recuperado de: https://www.redalyc.org/articulo.oa?id=377645733006.
28. Román García, (2011). Minería de datos en encuesta de profesores al fin de semestre de la Facultad de Ingeniería, UNAM. Recuperado de: https://tesiunam.dgb.unam.mx/F/YFFXKXAF2RJMU4BFUPJH3UFFGL1EYJUCEHU2C7SJS7A6SGYVFI-07205?func=full-set-set&set_number=027624&set_entry=000032&format=999

29. Solano, A., (2012). Toma de decisiones gerenciales. Tecnología en Marcha. Vol. 16 No 3, 44-51.
30. SOLARTE MARTINEZ, G. R., & OCAMPO S., C. A. (2009). TÉCNICAS DE CLASIFICACIÓN Y ANÁLISIS DE REPRESENTACION DEL CONOCIMIENTO PARA PROBLEMAS DE DIAGNÓSTICO. Scientia Et Technica, XV (42),177-182. ISSN: 0122-1701. Recuperado de: https://www.redalyc.org/articulo.oa?id=84916714033
31. UNAM, (2013). Filosofía, teoría y ciencia. Recuperado de: https://repositorio-uapa.cuaieed.unam.mx/repositorio/moodle/pluginfile.php/1470/mod_resource/content/2/contenido/index.html
32. Vélez Evans, M. I., (2006). El proceso de toma de decisiones como un espacio para el aprendizaje en las organizaciones. Revista Ciencias Estratégicas, 14(16),153-169. ISSN: 1794-8347. Recuperado de: https://www.redalyc.org/articulo.oa?id=151320326003.